1. (10 points) **(Who wins the election? Opinion Poll Strategy)** Imagine there are only two parties standing in the national election, and you have access to sampling and calling up uniformly random people from the electorate to find out who they're going to vote for. If the total population is N and an unknown $\frac{1}{4} \leq p \leq \frac{3}{4}$ fraction prefer BJP and $(1-p)$ prefer Congress, what's the maximum number of people you need to sample to estimate p upto an additive error of ϵ? Give the best possible answer up to constant factors, i.e., don't try to optimize the constants.

2. (10 points) **(More JL, More High-Dimensional Geometry)** There can only be d pairwise orthogonal vectors in $\mathbb{R}^d$ (as the span of these vectors is of dimension at least d). On the other hand, as we saw in HW2, for any $\epsilon > 0$, there can be exponentially (in the dimension d) many unit vectors which are nearly orthogonal in d dimensions, i.e., have absolute value of their pairwise inner products at most ϵ. Show that this is a consequence of the Johnson-Lindenstrauss Lemma.

   Theorem 1. *For any $\epsilon > 0$, and any integer d, there exists a collection of vectors $v_1, \ldots, v_t \in \mathbb{R}^d$, with $t = \exp(\Omega d)$ such that:*

   $$|\langle v_i, v_j \rangle| \leq \epsilon \qquad \forall \text{ distinct } i, j \in [t].$$

   Solution:

   Proof. $1 + 1 = 2$ □

3. (10 points) **(LP-Based Approach for P=NP?)** We will now attempt to show that a well-known NP-hard problem, the hamilton path problem is solvable using Linear Programming. Your goal is to find out if we are making any errors in our thought process. Indeed, our high level approach is to formulate the Hamilton Path problem as a flow problem, and solve the flow problem integrally.

   (a) (0 points) In class, we saw the flow LP polytope. Here is a similar LP formulation for the *minimum cost flow* problem. That is, we are given a directed graph $D = (V, A)$, each arc a has two integer capacities ℓ_a and u_a with $\ell_a \leq u_a$, and also a non-negative cost c_a. There is also a source vertex s and destination vertex t. The goal is to find the minimum cost way to send 1 unit flow from s to t in G such that the flow respects the upper and lower capacities on the different arcs.

Solution:

$$\begin{aligned}
\min & \sum_a c_a f_a \\
\text{s.t} & \sum_{a \in \delta_{\text{in}}(v)} f_a = \sum_{a \in \delta_{\text{out}}(v)} f_a && \forall v \in V \setminus \{s, t\} \\
& \sum_{a \in \delta_{\text{out}}(s)} f_a - \sum_{a \in \delta_{\text{in}}(s)} f_a = 1 \\
& \sum_{a \in \delta_{\text{in}}(t)} f_a - \sum_{a \in \delta_{\text{out}}(t)} f_a = 1 \\
& \ell_a \leq f_a \leq u_a && \forall a \in A
\end{aligned}$$

(b) (5 points) In general, the flow polytope above is integral if the lower and upper bounds are integral. That is, an optimal solution of the LP assigns flow values which are integer values as long as the capacities are all integer values also. Now, consider an undirected graph $G = (V, E)$ where each edge e has a cost 1, and a source vertex s and sink vertex t. Suppose each edge has lower and upper bounds on the flow as 0 and 1, and additionally each *vertex* now has a lower bound of flow as 1 and upper bound also as 1. Suppose the goal is to find the minimum cost flow in this problem. Show how to reduce this problem to one introduced in the previous part.

Solution:

Proof. $1 + 1 = 2$. □

(c) (5 points) Finally, we attempt to solve the following Hamilton path problem using the previous part: In this problem, given a graph G and s and t, the goal is to determine if there is an s–t path that visits each vertex exactly once. Suppose given this instance, we solve the flow problem defined in the previous part. Since our capacity bounds are integral, the resulting flow LP is also integral. Will this recover the Hamilton path? Explain what could go wrong if anything goes wrong.

Solution:

Proof. $1 + 1 = 2$. □

4. (15 points) **(Ambulance Migration)** In this question, you will design an algorithm for the following problem: The input consists of n locations, along with a metric space (represented by pairwise distances $d(i, j)$ between point i and point j which satisfy $d(i, j) + d(j, k) \geq d(i, k)$). Initially k ambulances are all in a central hospital at location

i_0. We are also given a sequence of locations i_1, i_2, ..., i_T which need to be serviced by these ambulances on successive days $1, 2, \ldots, T$. Your goal is to find out how to move ambulances to these locations so as to minimize the total movement cost. For example, if there are only two locations i_1 and i_2 (i.e., $T = 2$), there can only be two possible strateges: one strategy is to move one ambulance to location i_1 on day 1 and then the same ambulance to i_2 on day two. The cost here is $d(i_0, i_1)+d(i_1, i_2)$. Alternately, another strategy is to move one ambulance to i_1 on day 1, and move the second ambulance to i_2 on day 2 at a total cost of $d(i_0, i_1) + d(i_0, i_2)$. Can you design an algorithm with running time polynomial in n and T to compute the minimum cost movement sequence. Remember that the requests are ordered, that is, we can't service i_2 before i_1.

5. (15 points) **(Datastructure Design)** Suppose we have n points on the plane, $x_1, \ldots, x_n \in \mathbb{R}^k$ and wish to construct an algorithm that answers queries, given $y \in \mathbb{R}^k$, for maximizing and minimizing the inner product $\langle x_i, y \rangle, i \in [n]$. In other words, given y, the algorithm should output i and j that (respectively) maximize and minimize the inner product with y.

   (a) (10 points) Show that when $k = 2$ (that is, the points are on a plane), we can design algorithm that answer the queries in time $O(\log n)$.

   (b) (5 points) How do you extend the data-structure designed above for larger values of k. Assume k is a parameter much smaller than n and thus, we wish to minimize the running-time in terms of n first, and then k.

A Lecture Materials

A.1 Tail Inequalities

Often in our analysis, we model an interesting quantitity as a (real) random variable X and want to bound the tail of X (that is probability of X taking large values). The *Markov's inequality* states that when $X \geq 0$:

$$\mathbf{Pr}\left[X \geq t\,\mathbf{E}\left[X\right]\right] \leq 1/t.$$

Often, the variance,

$$\mathrm{Var}\left[X\right] := \mathbf{E}\left[(X - \mathbf{E}\,X)^2\right],$$

is known to be small. Applying the Markov's inequality to the square-deviation: $(X - \mathbf{E}[X])^2$, a non-zero random variable for any X, we have the *Chebyshev's Inequality*:

$$\mathbf{Pr}\left[|X - \mathbf{E}\left[X\right]| \geq t\,\mathrm{Var}\left[X\right]\right] \leq 1/t^2$$

A.2 Johnson-Lindenstrauss Lemma

The JL Lemma is a hallmark of dimension-reduction techniques.

Lemma 2. *For every $\epsilon > 0$, and every collection of m points, $x_1, \ldots, x_m \in \mathbb{R}^n$, there is a mapping of the points: $(x_j \to y_j)$, with $y_j \in \mathbb{R}^k$ for any $k \geq 8\log(m)/\epsilon^2$, where:*

$$(1-\epsilon)\cdot\left\|x_i - x_j\right\|_2 \leq \left\|y_i - y_j\right\|_2 \leq (1+\epsilon)\cdot\left\|x_i - x_j\right\|_2; \qquad \forall i,j \in [m].$$